UNITED STATES PATENT APPLICATION
FOR

# DEDICATED CACHE MEMORY

INVENTORS:

BLAISE B. FANNING,
a citizen of the United States

PREPARED BY:

BLAKELY, SOKOLOFF, TAYLOR & ZAFMAN LLP
12400 WILSHIRE BOULEVARD
SEVENTH FLOOR
LOS ANGELES, CA 90025-1030
(303) 740-1980

EXPRESS MAIL CERTIFICATE OF MAILING
"Express Mail" Mailing No. EV 331619579 US

I hereby certify that I am causing the above-referenced correspondence to be deposited with the United States Postal Service "Express Mail Post Office to Addressee" service on the date indicated below and that this paper or fee has been addressed to the Commissioner for Patents, Patent Application, P.O. Box 1450, Alexandria, VA 22313.
Date of Deposit: December 31, 2003
Name of Person Mailing Correspondence: Krista Mathieson

Krista Mathieson December 31, 2003
Signature Date

# DEDICATED CACHE MEMORY

## FIELD

[0001] An embodiment of the invention relates to computer storage in general, and more specifically to a dedicated cache memory.

## BACKGROUND

[0002] In computer operation, cache memories can improve system operation by providing access to certain data in memory that can be accessed more quickly than the mass storage. A processor and system may utilize multiple cache memories of differing sizes, locations, technologies, and operational speeds.

[0003] However, the use of cache memories may be more complicated when certain operations, such as when processes operate in allocated cycles. For example, a computer that receives a data stream, such as a multimedia stream, may pre-allocate certain compute cycles to the processing of the data stream to enable predictable and reliable operations.

[0004] If data processing for a particular operation is handled in certain allocated cycles, it is possible that a cache memory will be flushed in the intervals in between the allocated cycles. As a result, the data stream may not be able to utilize the cache to enable efficient operations because accesses to the cache will likely not result in usable cached data.

## BRIEF DESCRIPTION OF THE DRAWINGS

**[0005]** The invention may be best understood by referring to the following description and accompanying drawings that are used to illustrate embodiments of the invention. In the drawings:

**[0006]** **Figure 1** illustrates an embodiment of a microprocessor utilizing a dedicated cache memory;

**[0007]** **Figure 2** illustrates an embodiment of a cache memory operation;

**[0008]** **Figure 3** is a flow chart illustrating an embodiment of dedicated cache memory processes;

**[0009]** **Figure 4** is an illustration of an embodiment of dynamic establishment and modification of a dedicated cache memory; and

**[0010]** **Figure 5** illustrates an embodiment of a computer environment.

## DETAILED DESCRIPTION

**[0011]** A method and apparatus are described for a dedicated cache memory.

**[0012]** Before describing an exemplary environment in which various embodiments of the present invention may be implemented, some terms that will be used throughout this application will briefly be defined:

**[0013]** As used herein, "cache" or "cache memory" means a memory in which data is stored for fast retrieval. For example, a cache memory may comprise a small, fast memory to hold recently accessed data and to enable quicker subsequent access to the data. In one example, data that is read from or written to a main memory may be copied to a cache memory. In another example, data may be prefetched to store to a cache memory to enable more efficient operations.

**[0014]** As used herein, "thread" or "computing thread" means a part, path, task, or route of execution within a program, routine, or other process. A thread may be executed independently of other threads.

**[0015]** According to an embodiment of the invention, cache memory for a computer includes a dedicated cache memory. The dedicated cache memory is dedicated to one or more computer operations. Under one embodiment, the dedicated cache memory is a thread-specific memory that dedicates one or memories or sectors of memories for a certain thread or threads.

**[0016]** According to one embodiment of the invention, a dedicated cache is part of a cache memory, the cache memory comprising a general-purpose portion or sector and a dedicated portion or section. According to another embodiment of the

invention, a dedicated cache comprises a memory that is separate from a general-purpose cache memory.

**[0017]** Under an embodiment of the invention, the allocation of memory for a dedicated cache may be dynamic and may be created, modified, or eliminated as necessary. For example, a computer system encountering a process that would benefit from a dedicated cache may establish the dedicated cache.

**[0018]** An embodiment of the invention can be used for any computer process. An embodiment may be particularly useful for processes or threads that are active during certain allocated cycles or time slices. In such processes, data in a general-purpose cache may be flushed between allocated cycles. An embodiment of a dedicated cache may provide for a cache memory that is insulated from the operations of a general-purpose cache and thus may retain data for access by a particular thread or process without regard to operations in the general-purpose cache.

**[0019]** Under a particular embodiment of the invention, a dedicated cache memory is utilized for multi-media data. Multi-media operations may be extremely time sensitive and cache performance can greatly affect operational performance. An embodiment of the invention may be utilized to provide a thread-specific cache memory for a processor to enable predictable performance for multimedia encoding and decoding operations. For example, general-purpose PCs (personal computers) may be used to play DVD (digital versatile disk, previously referred to as digital video disk) or other media streams to a television out device. In future applications, computing platforms may be used to broadcast multiple streams of audio and video data over wired and wireless networks throughout a home or other location. For successful operation, a

general-purpose computing platform may be required to deliver accurate (or glitchless) media streams, even in the presence of other computing workloads such as word processors, spreadsheets, and Internet browsers. Under an embodiment of the invention, to enable processes to operate in a predictable and reliable manner, multimedia encoders and decoders that are provided access to pre-allocated compute cycles to allow timely processing of data streams are allocated thread-specific cache memory.

**[0020]** The time required to execute a predefined operation that requires a predefined number of calculations may vary substantially according to applicability of cached data to the relevant workload. Under an embodiment of the invention, a certain portion of cache memory is allocated the exclusive use of a particular thread or threads. In this manner, the data access time for a known workload may remain relatively constant and the amount of time required to process a worst-case data stream may be predicted.

**[0021]** In one example, computationally intensive workloads in current systems may require in excess of 1 gigabyte (GB) per second of main memory bandwidth. If it is assumed that a typical processor cache contains, for example, 1 megabyte (MB) of memory, then a simple division of these factors indicates that the entire cache may be flushed every millisecond. A multimedia thread that uses 10% of a system's computing power might be scheduled to compute for a 1-millisecond time slice every 10 milliseconds. In approximate terms, a cache memory may potentially be completely flushed nine times between activations of the multimedia thread.

**[0022]** Under an embodiment of the invention, a dedicated cache memory may be used by any operations that share the targeted thread ID. In the example of a thread that is scheduled to compute for a 1 millisecond time slice every 10 milliseconds, the

processor, although it executes general-purpose code for 9 milliseconds between media thread activations, does not disturb the section or sections of the cache memory dedicated for specific media operations. As a result, the media instruction streams are not required to waste allocated computes by reinitializing cache data.

**[0023]** **Figure 1** illustrates an embodiment of a microprocessor utilizing a dedicated cache memory. In this illustration, a processor **105** includes a processor core **110** for processing of operations and one or more cache memories. The cache memories may be structured in various different ways. Using common terminology for cache memories, the illustration shown in Figure 1 includes an L0 memory **115** that comprises a plurality of registers. Included on the processor **105** is an L1 cache **120** to provide very fast data access. Separate from the processor **105** is an L2 cache **130**, which generally will be larger but not as fast as the L1 cache **120**. A system may include other cache memories, such as the L3 cache **140** that is illustrated communicating with the processor through the L2 cache **130**.

**[0024]** In the illustration shown in Figure 1, one or more of the cache memories includes a portion or section that acts as a dedicated cache memory. For example, L1 cache **120** includes dedicated cache **125**, L2 cache **130** includes dedicated cache **135**, and L3 cache **140** includes dedicated cache **145**. Under another embodiment of the invention, the dedicated cache may be separate from the general-purpose cache memory. In one possible example, dedicated cache **155** may act in parallel with the L2 cache **130**. The dedicated cache memories are dedicated to certain processes, such as to data for certain threads.

[0025] **Figure 2** illustrates an embodiment of a cache memory operation. In this illustration, a cache memory **205** includes a general-purpose portion **210**, to be utilized by multiple operations, and a dedicated portion **215**. The dedicated portion **215** includes one or more sub-portions, with each sub-portion being a dedicated cache memory for a particular thread. In this illustration, the dedicated portion **215** includes a cache for a first thread designated as Thread 1 **220**, a cache for a second thread designated as Thread 2 **225**, and a cache for a third thread designated as Thread 3 **230**. Any data to be cached that relates to the operation of the relevant threads will be cached in the appropriate dedicated cache memory in the dedicated portion **215**. Any other data to be cached will be cached in the general-purpose portion **210**.

[0026] In one example, five different data elements are to be cached. For the purposes of this illustration, it is assumed that there are five possible threads, these threads being Thread 1 through Thread 5. A data stream **260** comprises data to be cached, with the data elements relating to various different threads. In this illustration, data for Thread 3 **235** is cached in the Thread 3 dedicated cache **230**. Data for Thread 5 **240** is not related to a dedicated cache and is therefore cached in the general-purpose portion **210** of cache memory **205**. Data for Thread 1 **245** is cached in the Thread 1 dedicated cache **220**. Data for Thread 2 **250** is cached in the Thread 2 dedicated cache **225**. Data for Thread 4 **255** is not related to a dedicated cache and is therefore cached in the general-purpose portion **210** of cache memory **205**.

[0027] **Figure 3** is a flow chart illustrating an embodiment of operation of a dedicated cache memory. For Figure 3, it is assumed that a system comprises a general-purpose cache and at least one dedicated cache. In this illustration, the dedicated

cache is dedicated to a specific thread. In Figure 3, a request to cache certain data is received **305**. For example, recently accessed data may be submitted to a cache. There is a determination regarding the thread ID for the data **310**. There is then a determination whether thread ID matches a thread-specific cache **315**. If so, operations are performed in the thread-specific cache **320**. If not, operations are performed in the general-purpose cache **325**.

**[0028]** **Figure 4** is an illustration of an embodiment of dynamic establishment and modification of a dedicated cache memory. Figure 4 illustrates certain operations that may be included in dynamic operations. In this illustration, a cache operation is performed **405**. There is a determination whether a dedicated cache for the appropriate thread exists **410**. If not, there is a determination whether a dedicated cache is needed **415**. If there is a need for a dedicated cache, a dedicated cache for the appropriated thread is allocated **420**.

**[0029]** If a thread-specific cache existed or has been created, there are then determinations regarding the size and continued existence of the cache. If a larger cache is needed **425**, the cache size may be increased **440**. If a smaller cache would be sufficient **435**, the cache size may be reduced **440**. If an existing dedicated cache for a particular thread is no longer required **445**, such as when the thread is no longer active, the dedicated cache may be eliminated **450**.

**[0030]** **Figure 5** illustrates an embodiment of an exemplary computer environment. Under an embodiment of the invention, a computer **500** comprises a bus **505** or other communication means for communicating information, and a processing

means such as one or more processors **510** (shown as **511** through **512**) coupled with the first bus **505** for processing information.

**[0031]** The computer **500** further comprises a random access memory (RAM) or other dynamic storage device as a main memory **515** for storing information and instructions to be executed by the processors **510**. Main memory **515** also may be used for storing temporary variables or other intermediate information during execution of instructions by the processors **510**. The computer **500** also may comprise a read only memory (ROM) **520** and/or other static storage device for storing static information and instructions for the processor **510**.

**[0032]** A data storage device **525** may also be coupled to the bus **505** of the computer **500** for storing information and instructions. The data storage device **525** may include a magnetic disk or optical disc and its corresponding drive, flash memory or other nonvolatile memory, or other memory device. Such elements may be combined together or may be separate components, and utilize parts of other elements of the computer **500**.

**[0033]** The computer **500** may also be coupled via the bus **505** to a display device **530**, such as a liquid crystal display (LCD) or other display technology, for displaying information to an end user. In some environments, the display device may be a touch-screen that is also utilized as at least a part of an input device. In some environments, display device **530** may be or may include an auditory device, such as a speaker for providing auditory information. An input device **540** may be coupled to the bus **505** for communicating information and/or command selections to the processor **510**. In various implementations, input device **540** may be a keyboard, a keypad, a touch-screen and stylus, a voice-activated system, or other input device, or combinations

of such devices. Another type of user input device that may be included is a cursor control device **545**, such as a mouse, a trackball, or cursor direction keys for communicating direction information and command selections to processor **510** and for controlling cursor movement on display device **530**.

**[0034]** A communication device **550** may also be coupled to the bus **505**. Depending upon the particular implementation, the communication device **550** may include a transceiver, a wireless modem, a network interface card, or other interface device. The computer **500** may be linked to a network or to other devices using the communication device **550**, which may include links to the Internet, a local area network, or another environment. In an embodiment of the invention, the communication device **550** may provide a link to a service provider over a network.

**[0035]** In the description above, for the purposes of explanation, numerous specific details are set forth in order to provide a thorough understanding of the present invention. It will be apparent, however, to one skilled in the art that the present invention may be practiced without some of these specific details. In other instances, well-known structures and devices are shown in block diagram form.

**[0036]** The present invention may include various processes. The processes of the present invention may be performed by hardware components or may be embodied in machine-executable instructions, which may be used to cause a general-purpose or special-purpose processor or logic circuits programmed with the instructions to perform the processes. Alternatively, the processes may be performed by a combination of hardware and software.

[0037] Portions of the present invention may be provided as a computer program product, which may include a machine-readable medium having stored thereon instructions, which may be used to program a computer (or other electronic devices) to perform a process according to the present invention. The machine-readable medium may include, but is not limited to, floppy diskettes, optical disks, CD-ROMs (compact disk read-only memory), and magneto-optical disks, ROMs (read-only memory), RAMs (random access memory), EPROMs (erasable programmable read-only memory), EEPROMs (electrically erasable programmable read-only memory), magnetic or optical cards, flash memory, or other type of media/machine-readable medium suitable for storing electronic instructions. Moreover, the present invention may also be downloaded as a computer program product, wherein the program may be transferred from a remote computer to a requesting computer by way of data signals embodied in a carrier wave or other propagation medium via a communication link (e.g., a modem or network connection).

[0038] Many of the methods are described in their most basic form, but processes can be added to or deleted from any of the methods and information can be added or subtracted from any of the described messages without departing from the basic scope of the present invention. It will be apparent to those skilled in the art that many further modifications and adaptations can be made. The particular embodiments are not provided to limit the invention but to illustrate it. The scope of the present invention is not to be determined by the specific examples provided above but only by the claims below.

[0039] It should also be appreciated that reference throughout this specification to "one embodiment" or "an embodiment" means that a particular feature may be included in the practice of the invention. Similarly, it should be appreciated that in the foregoing description of exemplary embodiments of the invention, various features of the invention are sometimes grouped together in a single embodiment, figure, or description thereof for the purpose of streamlining the disclosure and aiding in the understanding of one or more of the various inventive aspects. This method of disclosure, however, is not to be interpreted as reflecting an intention that the claimed invention requires more features than are expressly recited in each claim. Rather, as the following claims reflect, inventive aspects lie in less than all features of a single foregoing disclosed embodiment. Thus, the claims are hereby expressly incorporated into this description, with each claim standing on its own as a separate embodiment of this invention.